Cell Cycle

ISSN: 1538-4101 (Print) 1551-4005 (Online) Journal homepage: http://www.tandfonline.com/loi/kccy20

# DNA polymerase η mutational signatures are found in a variety of different types of cancer

Igor B. Rogozin, Alexander Goncearenco, Artem G. Lada, Subhajyoti De, German Nudelman, Anna R. Panchenko, David N. Cooper & Youri I. Pavlov

**To cite this article:** Igor B. Rogozin, Alexander Goncearenco, Artem G. Lada, Subhajyoti De, German Nudelman, Anna R. Panchenko, David N. Cooper & Youri I. Pavlov (2017): DNA polymerase η mutational signatures are found in a variety of different types of cancer, Cell Cycle, DOI: 10.1080/15384101.2017.1404208

**To link to this article:** https://doi.org/10.1080/15384101.2017.1404208

Accepted author version posted online: 15 Nov 2017.

**Publisher**: Taylor & Francis

**Journal**: *Cell Cycle*

**DOI**: https://doi.org/10.1080/15384101.2017.1404208



# DNA polymerase η mutational signatures are found in a variety of different types of cancer

Igor B. Rogozin[1], Alexander Goncearenco[1], Artem G. Lada[2], Subhajyoti De[3], German Nudelman[4], Anna R. Panchenko[1], David N. Cooper[5], Youri I. Pavlov[6,7]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, U.S.A.

[2] Department Microbiology and Molecular Genetics, University of California, Davis, CA, U.S.A.

[3] Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, U.S.A.

[4] Department of Neurology, Icahn School of Medicine at Mount Sinai; Systems Biology Center, Icahn School of Medicine at Mount Sinai, New York, New York 10029, U.S.A.

[5] Institute of Medical Genetics, School of Medicine, Cardiff University, UK

[6] Eppley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, Omaha, NE, U.S.A.

[7] Departments of Microbiology and Pathology; Biochemistry and Molecular Biology; Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, U.S.A.

**Short title:** DNA polymerase η in cancer cells

**Corresponding authors:**

Dr. Igor Rogozin, rogozin@ncbi.nlm.nih.gov

Dr. Youri Pavlov, ypavlov@unmc.edu

## Abstract

DNA polymerase (pol) η is a specialized error-prone polymerase with at least two quite different and contrasting cellular roles: to mitigate the genetic consequences of solar UV irradiation, and promote somatic hypermutation in the variable regions of immunoglobulin genes. Misregulation and mistargeting of pol η can compromise genome integrity. We explored whether the mutational signature of pol η could be found in datasets of human somatic mutations derived from normal and cancer cells. A substantial excess of single and tandem somatic mutations within known pol η mutable motifs was noted in skin cancer as well as in many other types of human cancer, suggesting that somatic mutations in A:T bases generated by DNA polymerase η are a common feature of tumorigenesis. Another peculiarity of pol ηmutational signatures, mutations in YCG motifs, led us to speculate that error-prone DNA synthesis opposite methylated CpG dinucleotides by misregulated pol η in tumors might constitute an additional mechanism of cytosine demethylation in this hypermutable dinucleotide.

**Key words:** hypermutation, *POLH*, mutable motif, DNA lesion bypass, sloppy DNA polymerase, skin cancer, gene expression profiles

## Introduction

The etiology of cancer lies in changes of genetic programming within the cell. Over the last decade, advances in sequencing technologies have potentiated the sequencing of whole genomes of both liquid and solid cancers (as well as individual tumor cells) giving birth to the new field of cancer genomics. One of the most significant discoveries has been that cancer genomes differ from the genomes of normal cells in their immediate vicinity in terms of thousands of newly acquired cancer-driving and passenger mutations [1-3], in perfect accordance to "mutator" theory of cancer [4, 5]. Multiple mutagenic processes, instigated by hereditary defects, or driven by intrinsic and environmental mutagens, contribute to this "genetic collapse" that changes the identity of cells [6-8]. The spectrum of genetic changes includes point mutations and other micro-lesions, chromosomal rearrangements and copy number changes that can be characteristic of both cancer and tissue type. For example, different types of tumor differ strikingly between mouse strains with defective exonucleases of pol δ versus pol ε [9] or when different members of APOBEC family are expressed, such as activation-induced deaminase AID (predominantly liquid tumors) versus APOBEC3B (breast and other solid tumors) in humans [10-12]. The hereditary lack of mismatch repair and/or exonuclease activity of replicative DNA pols predispose to colorectal cancer [13-15]; abnormal DNA double strand break repair leads to an increase in incidence of breast and ovarian cancer [16]; sunlight and defective pol η cause skin cancer [17].

Normal somatic cells also acquire mutations induced by the abovementioned plethora of factors during an individual's life time, albeit at lower rates than in tumors. For instance, comparison of the mutational burden in skin fibroblasts from forearm and hip from the same donors, revealed that the UV-induced (primarily C:G > T:A and CC:GG > TT:AA changes) and endogenous mutation rates per year in exposed skin were more than two-fold higher than that in unexposed areas [18]. This is in accord with previous studies of somatic mutations in sun-exposed skin [19, 20]. Cyclobutane pyrimidine dimers and (6-4)

photoproducts are the two major classes of lesion generated in DNA by UVB and UVC irradiation. Bypass of UV-induced photoproducts at TT tandem bases by the yeast and human translesion pol η (a member of the Y family of specialized DNA polymerases) is relatively accurate; this polymerase inserts the complementary AA nucleotides into the newly synthesized DNA in more than 99% of bypass events (measured using steady-state kinetic assays), thereby bypassing the lesion and suppressing the mutagenic effect of UV-induced DNA damage [21].

DNA pol η copies undamaged DNA with a lower fidelity than most DNA-directed polymerases with an average base-substitution error rate of $3.5 \times 10^{-2}$ [22-24]. Germline mutations in the gene (*POLH*) encoding DNA pol η result in XPV, a variant type of xeroderma pigmentosum [25]. Analysis of somatic mutations has suggested that transcription-coupled repair systems and DNA pol η are involved in the control of generation of somatic mutations in normal skin cells [18, 20]. It has also been noted that the 'pol η mutational signature' (Signature 9; http://cancer.sanger.ac.uk/cosmic/signatures) occurs in chronic lymphocytic leukemia and malignant B-cell lymphoma genomes [6, 26, 27]. "Signature 9" is characterized by a pattern of mutations that has been attributed to pol η (see the Discussion section for details) recruited for the repair of DNA damaged by AID during somatic hypermutation in immunoglobulin genes [28-30]. The mutable motif of pol η, the short motif WA/TW (W=A or T) was delineated in the context of somatic hypermutations and *in vitro* systems [22, 23]). We detected this signature in follicular lymphomas, but only significant in 5'UTR regions {Rogozin, 2016 #380, 29]. A recent study suggested that pol η may cause somatic mutations in lymphoid cells [31]; most of the characteristic clustered mutations were found in promoters, as with AID-initiated somatic hypermutation. In solid tumors, however, somatic mutations are likely to be associated with the other factors, including exogenous exposures, UV radiation or alcohol consumption [31].

In this paper, we have studied the possible involvement of pol η in the generation of somatic mutations in skin cancer, other cancers and in normal cells. A highly significant correlation between pol

η mutable motifs and somatic mutations in skin cancer cells was found. However, this correlation was not observed in normal skin samples. In addition to this, we also found traces of pol η mutagenesis in various other cancers. Taken together with the results of expression analysis, our study suggests the widespread participation of pol η in mutagenesis in cancer cells.

# Results

## Analysis of single and tandem somatic mutations found in normal skin samples

The starting point of our study was an analysis of single and tandem mutations in normal skin samples because of the known role of various DNA pols in the generation of somatic mutations in vigorously proliferating and exposed to environmental insults normal skin cells [9, 18, 20]. The majority of tandem double mutations are likely to be caused by the bypass of UV photoproducts formed between two pyrimidine residues, which is expected to be a significant feature of the mutational signature of pol η [32-34]. The dinucleotide mutabilities of CC, CT, TC and TT are actually strikingly different (Figure 1). TT dinucleotides have the lowest frequency of double and single mutations, consistent with the suggested antimutagenic property (see the Introduction section) of pol η while bypassing TT dimers. CC dinucleotides are extremely susceptible to changes (mostly transitions) and yielded the largest number of tandem double mutations (Figure 1).

Single mutations demonstrated a different propensity: the most frequently mutated are CG dinucleotides (Figure 1). It is well known that the motif YCG/CGR is hypermutated in human normal and cancer skin cells [18, 35]. CC dinucleotides were also found to be highly mutable although the frequency of mutation was lower than for CG dinucleotides (Figure 1). The third highest ranked mutable dinucleotide

was TC/GA. If we assume that pol η is responsible for the inaccurate bypass of dimers in CC, TC and CT dinucleotides, one would expect there to be an excess of single mutations in TC and CT dinucleotides (T is processed correctly and mutations arise while synthesizing past C nucleotides). We analyzed the excess of single mutations in TC/GA and CT/AG (positions of studied mutations are underlined). Examination of the DNA sequence context of mutations in these motifs showed that there was indeed a significant excess of substitutions (Table 1). The analysis was performed as described previously [36]. In brief, we calculated the excess of mutations in specific motifs using the ratio Fm/Fs, where Fm is the fraction of mutations observed in the particular motif, and Fs is the frequency of the motif in the respective DNA neighborhood (defined as a 120 bp DNA sequence window). A 1.2-fold excess of mutations (defined as described in Materials and Methods) in TC/GA and CT/AG dinucleotides was detected (Table 1). By contrast, there was no association between mutations and the WA/TW motif, associated with predominant errors of pol η when copying undamaged DNA [23, 29], indicating that pol η is unlikely to be involved in mutagenesis at undamaged DNA sites in normal skin cells (Table 1).

## Analysis of somatic mutations in skin cancer samples

Analysis of skin cancer cells strongly suggested that somatic mutations overlap with mutable motifs expected as a consequence of the error-prone bypass of photoproducts (TC/GA and CT/AG motifs) and the synthesis of undamaged DNA (WA/TW motifs) (Table 1). We also performed an analysis of two skin cancer subtypes with the highest representation in the COSMIC data set (see Materials and Methods), skin cutaneous melanoma and skin adenocarcinoma. A substantial (and significant) excess of somatic mutations for the both mutable motifs was found for skin cutaneous melanoma (Table 1) where the frequency of UV photoproducts is expected to be high. However, no such excess was found for skin adenocarcinoma (Table 1), consistent with the fact that adenocarcinoma initiates in the glandular cells

that are located deep inside or even under skin tissues, where no elevated frequency of UV photoproducts and mutations caused by DNA pol η in pyrimidine dinucleotides is to be expected.

## Analysis of somatic mutations in cancers other than skin

Previously, we found a signature of pol η (WA/TW) in follicular lymphoma which was significant only in 5'UTR regions (P-value = 0.01) [36]. Thus, it was suggested that a somatic mutational process operates in these regions in the "standard immunoglobulin mode" (significant correlation of mutation context with WRCH/DGYW and WA/TW mutable motifs, R = G/A, Y = C/T, D=A/T/G). The 5'UTR regions are known to be preferentially targeted by deaminases in actively transcribed genes [37, 38]. This is consistent with earlier studies that suggested that pol η may be mutagenic in chronic lymphocytic leukemia and malignant B-cell lymphoma genomes [39]. However, a more careful analysis of somatic mutations associated with pol η in follicular lymphoma suggests that this process is associated with translocations of the *BCL2* gene with immunoglobulin genes, a characteristic feature of follicular lymphoma [40]. Specifically, a detailed analysis of pol η mutability suggested that a substantial proportion (24%) of mutated 5'UTR WA/TW motifs occurred within the *BCL2* gene (19 out of 28 mutations at A:T bases). After we removed mutations that were identified within the *BCL2* 5'UTR region (near the translocation breakpoint), the correlation became insignificant (P-value = 0.11, 60 mutations in WA/TW motifs out of 116 mutations at A:T bases) [27]. This is one example of how a single mutation hotspot (in this case resulting from a translocation) is able to skew the results of the whole exome analysis, yielding misleading results.

Such discrepancies in results before and after elimination of somatic mutations associated with translocation events prompted us to analyze the pol η mutable motifs in different (sub)types of cancer. We did not find any significant excess of somatic mutations in WA/TW motifs in all types of blood cancer

merged together (Table 2). However, we did find such an excess in chronic lymphocytic leukemia and GCB lymphomas (subtypes of blood cancer) (Table 2), whereas no significant excess was found for acute myeloid leukemia (Table 2). This suggests that pol η may be mutagenic only in some types of blood cancer, consistent with the results of previous studies [39].

We found a significant excess of somatic mutations in WA/TW motifs in 11 out of 14 solid tumors from various tissue types (Table 2). Frequent tandem mutations are known to be an intrinsic property of DNA pol η when copying undamaged DNA and they have the same context specificity as single mutations [23]. Although tandem mutations occur much less frequently, we nevertheless found a significant excess of tandem mutations in the WA/TW context in 3 out of 8 cancer types (Table 3). A significant excess of tandem mutations in lung cancer (Table 3) appears to contradict the absence of any association between single somatic mutations and the WA/TW context (Table 2). This may result from greater sensitivity of the tandem mutation analysis or from differential representation of lung cancer subtypes in datasets of single and tandem mutations. To test the latter possibility, we performed an analysis of single mutations in two non-small cell lung cancer subtypes with the highest representation in the COSMIC data set, squamous cell carcinoma and adenocarcinoma. No significant association between mutations and the WA/TW context was found in lung adenocarcinoma, whereas a significant excess (1.3, P = 0.0005) of somatic mutations in the WA/TW context was found for lung squamous cell carcinoma suggesting that DNA pol η may be involved in mutagenesis in some lung cancer subtypes but not others. It should be noted that many (although not all) lung cancers are associated with cigarette smoking and exposure to a wide variety of exogenous mutagens, any of which could influence the observed mutational spectrum [6-8, 41].

Previously, we studied the role of AID in various cancer types and found the AID mutational signature to be prevalent in many types of human cancer, suggesting that AID-mediated, CpG

methylation-dependent mutagenesis is a common feature of tumorigenesis [36]. AID and DNA pol η are the two principal mutators involved in the somatic hypermutation of immunoglobulin genes that are coupled in the hypermutation machinery: AID is involved in the initiation of somatic hypermutation by massive cytosine deamination, whereas DNA pol η in involved in error-prone repair of DNA with the resulting lesions [28-30, 42]. We proposed to analyze the possible connection between these two enzymes in various cancer types using the excess of mutations in mutable motifs as independent variables. We found a negative correlation (CC = -0.44) between these two variables (Figure 2) which suggested that AID and DNA pol η are even decoupled in cancer-related mutagenesis (though the observed negative correlation is marginally significant, P = 0.044, one-tailed test).

## Analysis of somatic mutations in various normal tissues and expression analysis of Pol η

As a control, we examined the context of somatic mutations in various normal tissues [43] and did not find any significant excess of WA/TW mutable motifs (Supplemental Table 1). The size of these datasets is limited, but a power analysis (see Materials and Methods) suggested that the absence of any significant excess of somatic mutations in WA/TW mutable motifs in normal tissues likely reflects genuine biological properties of these samples.

We also compared the expression levels of the *POLH* gene (which encodes DNA pol η) in the various TCGA cohorts. Quartiles and extrema were calculated for each TCGA cohort selected in the study (Supplementary Figure 1). The observed high variability in *POLH* gene expression suggests that the gene is highly expressed only in a subset of TCGA tumor cohorts (Supplementary Figure 1) which is consistent with previous studies [44]. Specifically, *POLH* seems to be highly expressed in skin cutaneous melanoma (SKCM), consistent with a substantial and significant excess of pol η mutational signatures in this cancer

type (Table1). Previous analysis of an additional TCGA cohort with increased *POLH* expression, namely lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), suggested the possible involvement of pol η because of the presence of its characteristic mutation signature (Signature 9, Supplementary Figure 1, [39]). Notably, subsets of colorectal and uterine cancer (COAD, UCEC), which have been previously reported to have no association with polymerase Pol η activity, exhibit reduced *POLH* gene expression [45, 46].

## Discussion

A study of mutational signatures left by mutagenic enzymes, and, specifically, by pol η, can be augmented by investigating the expression profiles of the genes encoding for the enzymes in question [10, 47]. The TCGA atlas represents a comprehensive resource for the investigation of gene expression in the context of mutation datasets obtained from cohorts characterized by differing rates of somatic mutation. Observed heterogeneity in *POLH* gene expression within a comprehensive list of TCGA cohorts is consistent with previous reports suggesting that *POLH* activity is tissue- and tumor-specific [45]. Importantly, an elevated level of *POLH* expression was observed in tumor cohorts where pol η-specific mutation signatures were detected. Conversely, a reduced level of *POLH* expression was observed in tumor cohorts where no pol η-specific mutation signatures were detected.

The excess of pol η mutable motifs in chronic lymphocytic leukemia and GCB lymphomas that we detected in our work is consistent with the studies of Alexandrov et al. [39] where the pol η signature, "Signature 9", was detected in chronic lymphocytic leukemia and malignant B-cell lymphoma genomes. This is a promising result bearing in mind that the mutable motif of pol η (W<u>A</u>/<u>T</u>W) is rather short and hence less informative as compared to the AID/APOBEC mutable motifs [27, 29]. In general, "Signature 9" is

characterized by a pattern of mutations that has been widely attributed to pol η, although a higher frequency of T:A > G:C transversions compared to T:A > C:G transitions, although such a pattern has not been observed in studies of pol η either *in vitro* or *in vivo* [27, 29]. Although decomposition into signatures is a very useful tool for interpreting mutagenic processes, this approach has certain limitations [27]. One of them is the heuristic nature of the associations between mutational signatures and molecular mechanisms of mutation. In fact, we can never be sure that a given mutational signature can be attributed solely and exclusively to one molecular mechanism – indeed, some endogenous or exogenous mutational mechanisms may have very similar or even identical signatures [27, 29].

It should be stressed that important steps toward improving our understanding of the role of pol η in mutagenesis in skin cancer have been taken in previous studies, where the impact of transcription-coupled repair [20]and DNA pol η [18] in both normal and cancer skin cells were postulated. It should be noted that the strand-specificity (a signature of transcription-coupled repair) of mutations induced by pol η is well known in the context of the somatic hypermutation of immunoglobulin genes [18, 20, 48, 49]. Thus, all these studies point to pol η being an important mutagenic factor in normal skin and cancer cells. The recent study extended a range of potential mutagenic activity of pol η to solid tumors where somatic mutations produced by pol η are likely to be associated with the other factors, including exogenous exposures, UV radiation or alcohol consumption [31].

We detected overlaps between pol η signatures with somatic mutations in various cancers. It is possible that the perturbed cell metabolism leads to the aberrant regulation of pol η, for example, that what one expects in the completely disorganized environment of cell extracts (where the pol η mutagenesis had beed inferred) [50], or in *in vitro* systems [22], while normal cells are well protected from its action [51]. The error-prone action of misregulated pol η is expected to cause a substantial load of somatic mutations, which may be beneficial for cancer initiation and/or progression, for example, when TP53 is

mutated [27, 52]. Another potential function of pol η in cancer cells could be the error-free or error-prone bypass of various DNA lesions. It was suggested in a recent paper [53] that NPM1 (nucleophosmin) regulates translesion DNA synthesis (TLS) via an interaction with the catalytic core of pol η. NPM1 deficiency causes a TLS defect due to the proteasomal degradation of pol η. The prevalent *NPM1* mutation (c+) leading in one-third of AML patients to NPM1 mislocalization results in a loss of pol η, which may explain why no significant excess of mutation in pol η motifs was found for acute myeloid leukemia (Table 2). These results hint at the complexity of regulation of pol η in cancer cells and provide an explanation of why pol η mutational signatures are found only in some cancer types/subtypes.

It was suggested that, in both normal and cancer skin cells, a significantly increased frequency of UVB-induced transition mutations at YCG motifs could be explained by the participation of pol η [35]. Taking into account the high frequency of mutations in TC dinucleotides, it is tempting to speculate that mutagenesis of YCG motifs is caused by the error-prone synthesis by pol η on methylated cytosine in TCG/CCG sequences (with or without neighboring photoproducts) (Figure 1). Thus, the error-prone synthesis in YCG motif might be an additional mechanism of demethylation, by pol η misincorporating A instead of G in various types of cancer cell. The evidence for that comes from the observed negative correlation between the excess of somatic mutations associated with AID and pol η mutable motifs in various types of cancer (Figure 2). However, this hypothesis requires further experimental validation and would require analysis of methylated templates using *in vitro* pol η mutagenesis systems. The analysis of mammalian model species and cell cultures might also provide the means to test this hypothesis.

# Materials and Methods

## Analysis of somatic mutations

DNA sequences surrounding the mutated nucleotide represent the mutation context. We compared the frequencies of known mutable motifs for somatic mutations with the frequencies of these motifs in the vicinity of the mutated nucleotides. Specifically, for each base substitution, the 120 bp sequence centered around the mutation was extracted (the DNA neighborhood). We used only the nucleotides immediately surrounding the mutations because DNA pol η is thought to scan a limited length of DNA to mutate nucleotides in a preferred motif [36, 54]. This approach does not exclude any given region of the genome in general, but rather uses the areas within each sample where mutagenesis has happened (taking into account the variability in mutation rates across the human genome), and then evaluates whether the mutagenesis in this sample was enriched for DNA pol η motifs [36, 54]. This approach was thoroughly tested and its high accuracy demonstrated [36]. The frequencies of mutable motifs in the locations of somatic mutations was compared to the frequencies of the same motifs in the DNA neighborhood (Figure 3) using Fisher's exact test (2 x 2 table, 2-tail test) as previously described [36, 54] (for details see Figure 3).

The exome sequencing data of somatic mutations in normal skin cells were obtained from [20]. Somatic mutation data from the ICGC and TCGA cancer genome projects were extracted from the Sanger COSMIC Whole Genome Project v75 http://cancer-beta.sanger.ac.uk/cosmic. The tissues and cancer types were defined according to primary tumor site and cancer genome sequencing projects. Somatic mutations in various normal tissues were from Yadav *et al.* [43].

## Power analysis of mutations in normal tissues

We compared the magnitude of the difference between the fraction of mutations observed in the mutable motif and the fraction of motifs in the surrounding region (effect size) for somatic mutations in normal tissues. For the purpose of this comparison (power analysis), we used a sampling procedure that was repeated 1,000 times. Each sample of all available somatic mutations from cancer cells (where a significant excess of somatic mutations in WA/TW motifs was observed, the last row in the Table 2) had a size equal to that for normal tissues (552 somatic mutations, the last row in the Supplementary Table 1). Analysis of the difference between the fractions showed that the difference for normal mutations was smaller for 99.1% of cancer samples. Thus, the observed effect size (Supplementary Table 1) is likely to reflect the biological properties of these samples and is unlikely to result from the small sample size, at least for somatic mutations from normal tissues.

## Expression analysis of the POLH gene

For the *POLH* gene expression analysis, the normalized version of the RSEM (Broad Institute TCGA Genome Data Analysis Center. 2016, Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard Dataset http://doi.org/10.7908/C11G0KM9) was used to analyze the TCGA RNA-Seq datasets from the Broad Genome Data Analysis Center. For each TCGA cohort (Supplementary Figure 1), the low and upper bounds, median, outliers, and first and third quartiles were retrieved via the FireBrowse RESTful API (http://firebrowse.org/api-docs/) for the tumor and the corresponding normal (when available) tissue samples.

## Author contributions

IBR, DNC and YIP incepted the study; IBR, AG and AGL designed and performed data analysis; IBR, AGL, AG, ARP, MLG, SD, GN, DNC and YIP analyzed and interpreted the results; IBR, DNC and YIP wrote the manuscript that was edited and approved by all authors.

## Competing financial interests

The author(s) declare no competing financial interests.

## Acknowledgements

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health [to I.B.R., A.G., A.R.P.]; Boettcher Foundation, American Cancer Society, P30 CA072720 [to S.D.]; NE DHHS LB506, grant 2017-48 [to Y.I.P.]; the Fred & Pamela Buffett Cancer Center Support Grant from the National Cancer Institute under award number P30 CA072720 (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health) and Qiagen Inc through a License Agreement with Cardiff University [to D.N.C.].

**References**


1. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. Nat Rev Genet 2013; 14:703-18.

2. Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 2015; 518:495-501.

3. Wood LD, Hruban RH. Genomic landscapes of pancreatic neoplasia. J Pathol Transl Med 2015; 49:13-22.

4. Loeb LA, Springgate CF, Battula N. Errors in DNA replication as a basis of malignant changes. Cancer Res 1974; 34:2311-21.

5. Loeb LA. Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences. Cancer Res 2016; 76:2057-9.

6. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer. Nature 2013; 500:415-21.

7. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer 2014; 14:786-800.

8. Pfeifer GP. How the environment shapes cancer genomes. Current Opinion in Oncology 2015; 27:71-7.

9. Albertson TM, Ogawa M, Bugni JM, Hays LE, Chen Y, Wang Y, Treuting PM, Heddle JA, Goldsby RE, Preston BD. DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice. Proc Natl Acad Sci U S A 2009; 106:17101-4.

10. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 2013; 494:366-70.

11. Heintel D, Kroemer E, Kienle D, Schwarzinger I, Gleiss A, Schwarzmeier J, Marculescu R, Le T, Mannhalter C, Gaiger A, et al. High expression of activation-induced cytidine deaminase (AID) mRNA is associated with unmutated IGVH gene status and unfavourable cytogenetic aberrations in patients with chronic lymphocytic leukaemia. Leukemia 2004; 18:756-62.

12. Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon KR, Resch W, Liang G, Tang Z, Mathe E, Benner C, et al. B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. Cell 2014; 159:1524-37.

13. Fishel R, Kolodner RD. Identification of mismatch repair genes and their role in the development of cancer. Current Opinion in Genetics & Development 1995; 5:382-95.

14. Heitzer E, Tomlinson I. Replicative DNA polymerase mutations in cancer. Current Opinion in Genetics & Development 2014; 24:107-13.

15. Barbari SR, Shcherbakova PV. Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy. DNA Repair (Amst) 2017; 56:16-25.

16. Scully R. Role of BRCA gene dysfunction in breast and ovarian cancer predisposition. Breast Cancer Research : BCR 2000; 2:324-30.

17. Masutani C, Kusumoto R, Yamada A, Dohmae N, Yokoi M, Yuasa M, Araki M, Iwai S, Takio K, Hanaoka F. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. Nature 1999; 399:700-4.

18. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, Fargo DC, Boyer JC, Kaufmann WK, Taylor JA, et al. The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLoS Genet 2016; 12:e1006385.

19. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 2010; 107:961-8.

20. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science (New York, NY) 2015; 348:880-6.

21. Washington MT, Johnson RE, Prakash L, Prakash S. Accuracy of lesion bypass by yeast and human DNA polymerase eta. Proc Natl Acad Sci U S A 2001; 98:8355-60.

22. Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. Low fidelity DNA synthesis by human DNA polymerase h. Nature 2000; 404:1011-3.

23. Matsuda T, Bebenek K, Masutani C, Rogozin IB, Hanaoka F, Kunkel TA. Error rate and specificity of human and murine DNA polymerase eta. Journal of Molecular Biology 2001; 312:335-46.

24. Pavlov YI, Shcherbakova PV, Rogozin IB. Roles of DNA polymerases in replication, repair, and recombination in eukaryotes. International Review of Cytology 2006; 255:41-132.

25. Johnson RE, Kondratick CM, Prakash S, Prakash L. hRAD30 mutations in the variant form of xeroderma pigmentosum. Science (New York, NY) 1999; 285:263-5.

26. Petljak M, Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis 2016; 37:531-40.

27. Rogozin IB, Pavlov YI, Goncearenco A, De S, Lada AG, Poliakov E, Panchenko AR, Cooper DN. Mutational signatures and mutable motifs in cancer genomes. Briefings in Bioinformatics 2017.

28. Zanotti KJ, Gearhart PJ. Antibody diversification caused by disrupted mismatch repair and promiscuous DNA polymerases. DNA Repair (Amst) 2016; 38:110-6.

29. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nature Immunology 2001; 2:530-6.

30. Zeng X, Winter DB, Kasmer C, Kraemer KH, Lehmann AR, Gearhart PJ. DNA polymerase eta is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. Nature Immunology 2001; 2:537-41.

31. Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell 2017; 170:534-47.e23.

32. McCulloch SD, Kokoska RJ, Masutani C, Iwai S, Hanaoka F, Kunkel TA. Preferential cis-syn thymine dimer bypass by DNA polymerase eta occurs with biased fidelity. Nature 2004; 428:97-100.

33. Saribasak H, Maul RW, Cao Z, Yang WW, Schenten D, Kracker S, Gearhart PJ. DNA polymerase zeta generates tandem mutations in immunoglobulin variable regions. J Exp Med 2012; 209:1075-81.

34. Maul RW, MacCarthy T, Frank EG, Donigan KA, McLenigan MP, Yang W, Saribasak H, Huston DE, Lange SS, Woodgate R, et al. DNA polymerase iota functions in the generation of tandem mutations during somatic hypermutation of antibody genes. J Exp Med 2016; 213:1675-83.

35. Lee DH, Pfeifer GP. Deamination of 5-methylcytosines within cyclobutane pyrimidine dimers is an important component of UVB mutagenesis. J Biol Chem 2003; 278:10314-21.

36. Rogozin IB, Lada AG, Goncearenco A, Green MR, De S, Nudelman G, Panchenko AR, Koonin EV, Pavlov YI. Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. Scientific Reports 2016; 6:38133.

37. Lada AG, Kliver SF, Dhar A, Polev DE, Masharsky AE, Rogozin IB, Pavlov YI. Disruption of Transcriptional Coactivator Sub1 Leads to Genome-Wide Re-distribution of Clustered Mutations Induced by APOBEC in Active Yeast Genes. PLoS Genet 2015; 11:e1005217.

38. Taylor BJ, Wu YL, Rada C. Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. Elife 2014; 3:e03553.

39. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Current Opinion in Genetics & Development 2014; 24:52-60.

40. Green MR, Kihira S, Liu CL, Nair RV, Salari R, Gentles AJ, Irish J, Stehr H, Vicente-Duenas C, Romero-Camarero I, et al. Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. Proc Natl Acad Sci U S A 2015; 112:E1116-25.

41. Temiz NA, Donohue DE, Bacolla A, Vasquez KM, Cooper DN, Mudunuri U, Ivanic J, Cer RZ, Yi M, Stephens RM, et al. The somatic autosomal mutation matrix in cancer genomes. Human Genetics 2015; 134:851-64.

42. Laffleur B, Denis-Lagache N, Peron S, Sirac C, Moreau J, Cogne M. AID-induced remodeling of immunoglobulin genes and B cell fate. Oncotarget 2014; 5:1118-31.

43. Yadav VK, DeGregori J, De S. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. Nucleic Acids Research 2016; 44:2075-84.

44. Makridakis NM, Reichardt JK. Translesion DNA polymerases and cancer. Front Genet 2012; 3:174.

45. CGAN. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487:330-7.

46. CGARN. Integrated genomic analyses of ovarian carcinoma. Nature 2011; 474:609-15.

47. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet 2013; 45:970-6.

48. Pavlov YI, Rogozin IB, Galkin AP, Aksenova AY, Hanaoka F, Rada C, Kunkel TA. Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase eta during copying of a mouse immunoglobulin kappa light chain transgene. Proc Natl Acad Sci U S A 2002; 99:9954-9.

49. Mayorov VI, Rogozin IB, Adkison LR, Gearhart PJ. DNA polymerase eta contributes to strand bias of mutations of A versus T in immunoglobulin genes. Journal of immunology 2005; 174:7781-6.

50. Bebenek K, Matsuda T, Masutani C, Hanaoka F, Kunkel TA. Proofreading of DNA polymerase eta-dependent replication errors. J Biol Chem 2001; 276:2317-20.

51. Pavlov YI, Nguyen D, Kunkel TA. Mutator effects of overproducing DNA polymerase eta (Rad30) and its catalytically inactive variant in yeast. Mutat Res 2001; 478:129-39.

52. Lerner LK, Francisco G, Soltys DT, Rocha CR, Quinet A, Vessoni AT, Castro LP, David TI, Bustos SO, Strauss BE, et al. Predominant role of DNA polymerase eta and p53-dependent translesion synthesis in the survival of ultraviolet-irradiated human cells. Nucleic Acids Research 2017; 45:1270-80.

53. Ziv O, Zeisel A, Mirlas-Neisberg N, Swain U, Nevo R, Ben-Chetrit N, Martelli MP, Rossi R, Schiesser S, Canman CE, et al. Identification of novel DNA-damage tolerance genes reveals regulation of translesion DNA synthesis by nucleophosmin. Nat Commun 2014; 5:5437.

54. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, Kim J, Kwiatkowski DJ, Fargo DC, Mieczkowski PA, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet 2015; 47:1067-72.

**Legends to Figures**

Fig. 1

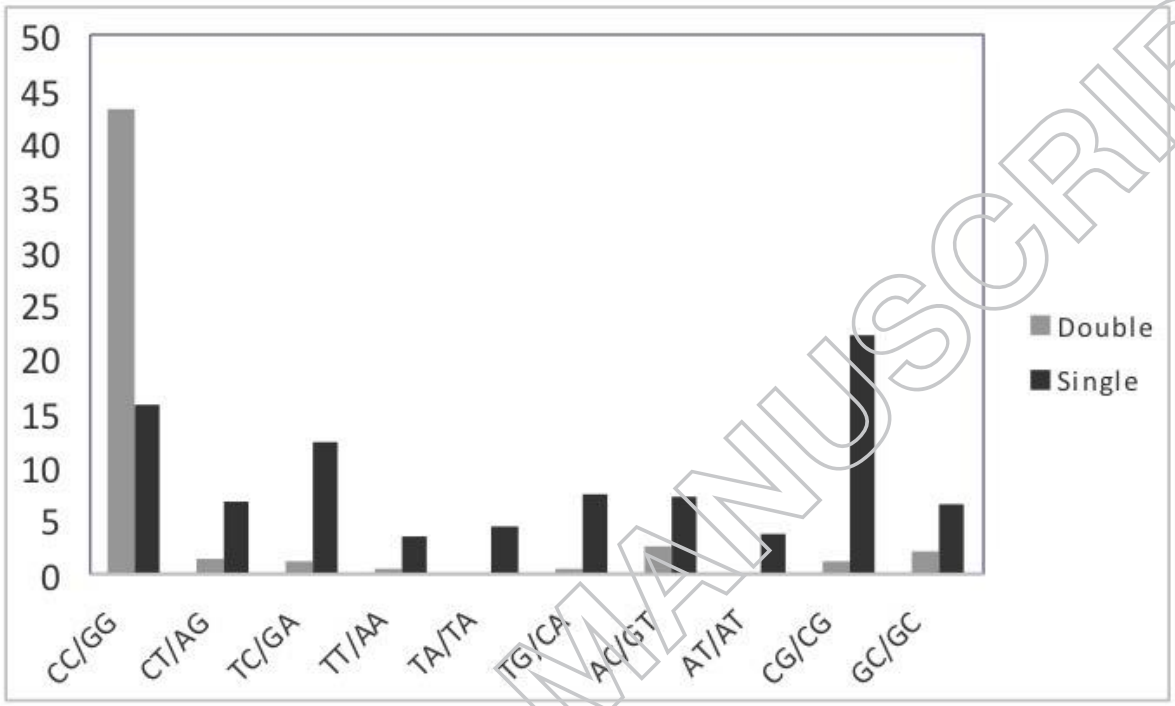


**Figure 1. Frequency of tandem double (blue) and single mutations (red) in various dinucleotides.** The Fnorm is a normalized frequency of double or single mutations (the number of mutations in dinucleotides XX multiplied by 1000 and divided by the number of dinucleotides XX in the DNA neighborhood)

Fig. 2

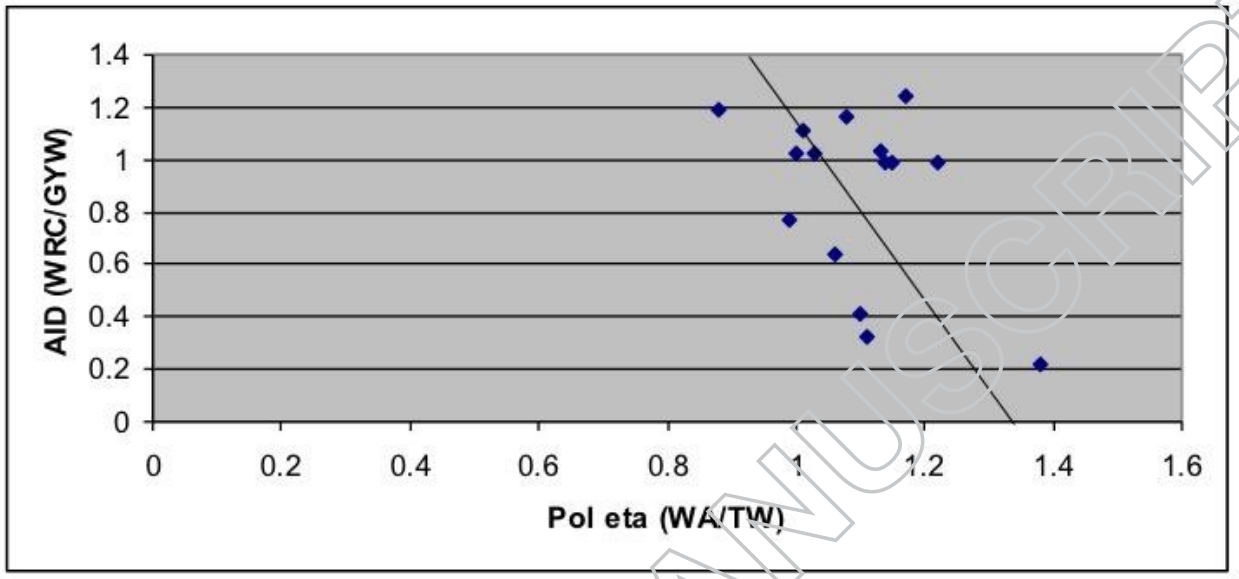


**Figure 2. Comparison of excess of somatic mutations associated with AID and pol η mutable motifs in various types of cancer.** The excess of mutations in motifs was calculated using the ratio Fm/Fs, where Fm is the fraction of tandem somatic mutations (both positions are used for this analysis) observed in the studied mutable motif (the number of mutated motifs divided by the number of mutations), and Fs is the frequency of the motif in the DNA context of somatic mutations (the number of motif positions divided by the total number of all un-mutated positions in surrounding regions). Linear correlation coefficient is –0.44 (P = 0.044, one-tail test). The regression line is shown in black.

Fig. 3

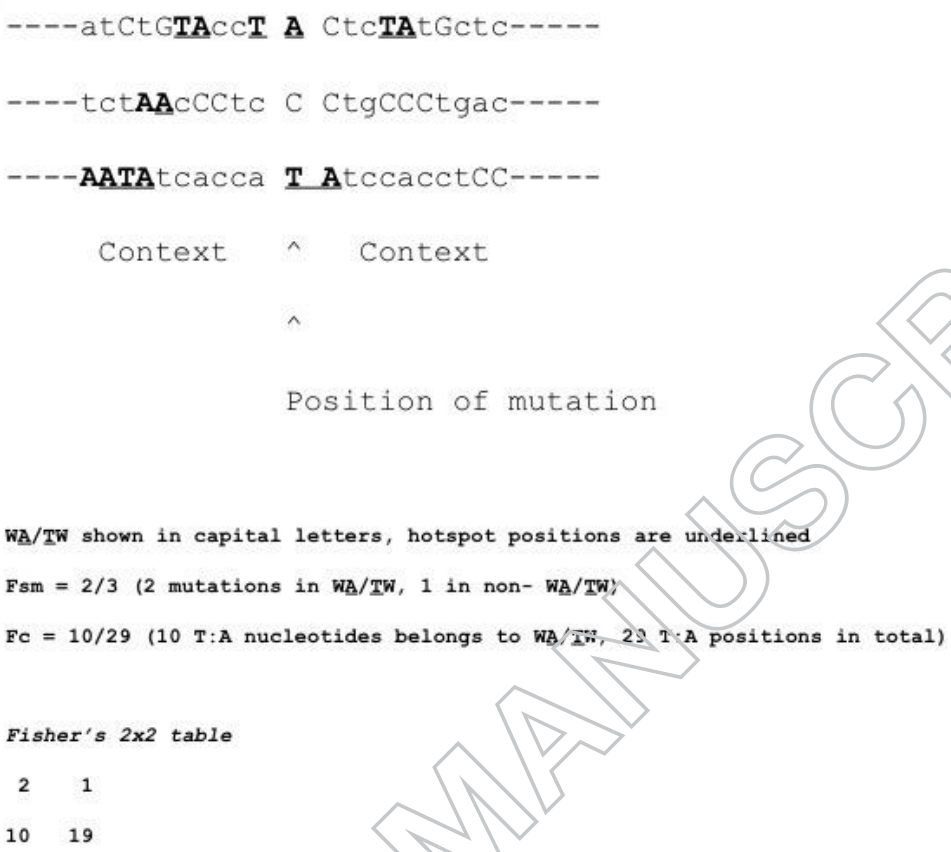


**Figure 3. Statistical analysis of mutable motifs in sites of somatic mutations and surrounding regions.** The excess of mutations in motifs was calculated using the ratio Fm/Fs, where Fm is the fraction of somatic mutations observed in the given mutable motif (the number of mutated motifs divided by the number of mutations), and Fs is the frequency of the motif in the DNA neighborhood of somatic mutations (the number of motif positions divided by the total number of all un-mutated positions in the 120 bp window).

Table 1. Association between DNA polymerase η mutable motifs (WA/TW)* and the DNA sequence context of somatic mutations in normal and cancer skin cells

| Mutable motif | Fraction of mutations observed in the mutable motif (Fm) vs. Fraction of motifs in surrounding regions (Fs) | Excess of mutations in the motif | P-value, Fisher's exact test ** |
|---|---|---|---|
| Normal skin cells | | | |
| TC/GA and CT/AG | 0.542. vs. 0.45 | 1.2 | $<10^{-10}$ |
| WA/TW | 0.435 vs. 0.424 | 1.03 | NS |
| Skin cancer (the Sanger COSMIC Whole Genome Project) | | | |
| TC/GA and CT/AG | 0.502 vs. 0.461 | 1.09 | $<10^{-10}$ |
| WA/TW | 0.593 vs. 0.432 | 1.37 | $<10^{-10}$ |
| Skin cancer subtypes: skin cutaneous melanoma | | | |
| TC/GA and CT/AG | 0.7 vs. 0.476 | 1.47 | $<10^{-10}$ |
| WA/TW | 0.6 vs. 0.422 | 1.42 | $<10^{-10}$ |
| **Skin cancer subtypes: skin adenocarcinoma** | | | |
| TC/GA and CT/AG | 0.406 vs. 0.427 | 0.95 | NS |

| | | | |
|---|---|---|---|
| W<u>A</u>/<u>T</u>W | 0.28 vs. 0.35 | 0.80 | NS |

* The correlation was measured using Fisher's exact test. Mutable positions in consensus sequences are underlined (W = A or T). The excess of mutations in motifs was calculated using the ratio Fm/Fs, where Fm is the fraction of somatic mutations observed in the given mutable motif (the number of mutated motifs divided by the number of mutations), and Fs is the frequency of the motif in the DNA neighborhood of somatic mutations (the number of motif positions divided by the total number of all un-mutated positions in the 120 bp window).

** NS, no significant excess

Table 2. Preferential mutability of DNA polymerase η mutable motifs (WA/TW) in various cancers (single mutations from Whole Genomes and Whole Exomes, the Sanger COSMIC Whole Genome Project)

| Tissue | Fraction of mutations observed in the mutable motif (total number of sites) | Fraction of motifs in surrounding regions (total number of sites) | Excess of mutations in the motif | P-value, Fisher's exact test* |
|---|---|---|---|---|
| Blood | 0.328 (8,269) | 0.372 (437,552) | 0.88 | NS |
| Chronic lymphocytic leukemia | 0.529 (412) | 0.435 (23,680) | 1.22 | 0.00009 |
| Acute myeloid leukemia | 0.29 (6,727) | 0.351 (348,871) | 0.83 | NS |
| GCB lymphomas | 0.49 (1,070) | 0.43 (61,426) | 1.44 | 0.00003 |
| Bladder | 0.468 (5,952) | 0.426 (339,359) | 1.1 | $<10^{-10}$ |
| Breast | 0.453 | 0.428 | 1.06 | $<10^{-10}$ |

| | | | | |
|---|---|---|---|---|
| | (18,453) | (1,068,627) | | |
| Cervix | 0.499 | 0.448 | 1.11 | $<10^{-10}$ |
| | (3,193) | (186,165) | | |
| Colon | 0.466 | 0.43 | 1.08 | $<10^{-10}$ |
| | (45,103) | (2,595,315) | | |
| Kidney | 0.482 | 0.424 | 1.14 | $<10^{-10}$ |
| | (19,290) | (1,113,567) | | |
| Liver | 0.424 | 0.426 | 1.0 | NS |
| | (44,028) | (2,520,549) | | |
| Lung | 0.419 | 0.422 | 0.99 | NS |
| | (45,264) | (2,592,238) | | |
| Ovary | 0.441 | 0.423 | 1.03 | 0.0006 |
| | (8,114) | (461,545) | | |
| Pancreas | 0.482 | 0.427 | 1.13 | $<10^{-10}$ |
| | (9,394) | (535,889) | | |
| Prostate | 0.493 | 0.43 | 1.15 | $<10^{-10}$ |
| | (13,036) | (775,226) | | |

| | | | | |
|---|---|---|---|---|
| Rectum | 0.537 | 0.441 | 1.22 | $<10^{-10}$ |
| | (8,213) | (482,509) | | |
| Skin | 0.593 | 0.430 | 1.38 | $<10^{-10}$ |
| | (26,859) | (1,541,263) | | |
| Stomach | 0.504 | 0.431 | 1.17 | $<10^{-10}$ |
| | (50,212) | (2,897,221) | | |
| Uterus | 0.440 | 0.438 | 1.01 | NS |
| | (55,999) | (3,212,849) | | |

Tissue types with significant correlation (taking into account the Bonferroni correction for multiple tests) between the motif and somatic mutations are underlined. The excess of mutations in motifs was calculated using the ratio Fm/Fs, where Fm is the fraction of somatic mutations observed in the studied mutable motif (the number of mutated motifs divided by the number of mutations), and Fs is the frequency of the motif in the DNA context of somatic mutations (the number of motif positions divided by the total number of all un-mutated positions in surrounding regions).

* Absence of significant excess of mutations in WA/TW (NS, no significant excess) suggests that there is no connection between mutagenesis and WA/TW motifs.

Table 3. Analysis of tandem somatic mutations in DNA polymerase η mutable motifs (WA/TW) in various cancers (Whole Genomes and Whole Exomes, the Sanger COSMIC Whole Genome Project)

| Tissue | Fraction of mutations observed in the mutable motif (total number of sites) | Fraction of motifs in surrounding regions (total number of sites) | Excess of mutations in the motif | P-value, Fisher's exact test* |
|---|---|---|---|---|
| Cervix | 0.66 (9) | 0.444 (960) | 1.49 | NS |
| Colon | 0.8 (5) | 0.4 (558) | 2. | NS |
| Kidney | 0.571 (7) | 0.286 (766) | 2. | NS |
| Lung | 1 (13) | 0.39 (1,322) | 2.6 | $5.3\times10^{-6}$ |
| Ovary | 0.944 (18) | 0.167 (1,806) | 5.65 | $4.7\times10^{-6}$ |
| Pancreas | 1 (8) | 0.547 (1,032) | 1.83 | NS |

| | | | | |
|---|---|---|---|---|
| Rectum | 1 | 0.244 | 4.1 | NS |
| | (3) | (262) | | |
| Skin | 0.847 | 0.576 | 1.47 | $6\times10^{-8}$ |
| | (59) | (6,953) | | |
| All somatic mutations | 0.858 | 0.47 | 1.83 | $<10^{-10}$ |
| | (134) | (15,097) | | |

Tissue types with significant correlation (taking into account the Bonferroni correction for multiple tests) between the motif and somatic mutations are underlined. The excess of tandem mutations in motifs was calculated using the ratio Fm/Fs, where Fm is the fraction of tandem somatic mutations (both positions are used for this analysis) observed in the studied mutable motif (the number of mutated motifs divided by the number of tandem mutations), and Fs is the frequency of the motif in the DNA context of tandem somatic mutations (the number of motif positions divided by the total number of all un-mutated positions in surrounding regions).

* Absence of significant excess of mutations in WA/TW (NS, no significant excess) suggests that there is no association between mutagenesis and WA/TW motifs.